

Dr. J. Lederberg

~~Dr. J. Lederberg~~

CITATION INDEX TO GENETICS AND GENERAL SCIENCE LITERATURE

RESEARCH PROPOSAL BY THE

INSTITUTE FOR SCIENTIFIC INFORMATION
PHILADELPHIA 23, PA.

BACKGROUND

During the past ten years there have been numerous published and unpublished expressions of need for a system of organizing citations in scientific literature. Seidel (1) and Hart (2) were familiar with the indispensability of citation indexes for legal searches. As patent attorneys they recognized the potential value of citation indexes for literature searching.(3) In 1955 Adair(4) and Garfield(5)(6) proposed citation indexes for science literature. In 1957 the National Science Foundation reported that their fiscal research program on scientific information would include studies on a "proposed method of bringing related material together similar in some respects to Shepard's Citations(7), a respected method in the field of law, which has never been tried in the sciences."(8) Following a suggestion from Berry of NSF, (9) a research proposal was then submitted to the NSF by Eugene Garfield Associates.(10) This proposal was rejected by NSF with a recommendation that it be resubmitted, with revisions, provided a group of recognized scientists in one branch of science advise and guide the course of this research.(11) Since Allen(12) and Lederberg(13) independently expressed interest in a science citation index (SCI) the field of genetics seemed a logical point of concentration. Further discussions took place with geneticists at the NSF(14) and NIH(15) as well as with the NIH genetics study section. An Advisory Committee to the SCI was subsequently formed consisting of:

Dr. Gordon Allen, National Institute of Mental Health
Dr. Joshua Lederberg, Stanford University
Dr. George LeFevre, Harvard University
Dr. Joseph Melnick, Baylor University
Dr. Sol Spiegelman, Univ. of Illinois

The following proposal is the result of considerable discussion with members of this board and interested persons at the NSF and NIH.

WHAT IS A CITATION INDEX

A citation index is a bibliography in which one finds citations to specific journal articles. These citations come from other journal articles that have appeared subsequent to the article cited.

HOW IS A CITATION INDEX USED

In the course of one's reading an article is found which is of interest. However, it may be several years old and one would like to be brought up-to-date on subsequent developments. Using the SCI one locates the citation for this article and there finds a list of all subsequent articles that have referred to it.

OF WHAT USE IS A SCIENCE CITATION INDEX(SCI)

There are many uses for the SCI. Most of them are analagous to their use in legal research. In Shepard's Citations one is able to quickly trace the history of a particular legal decision--whether it has been sustained one or more times, modified, or overruled. Similarly, in the SCI one can be brought up-to-date on a particular paper found in the literature. For example, if a new biochemical "method" is discovered one can quickly learn what other compounds, organisms, etc. have been prepared by the same method. If a statistical technique is employed by one author and adopted by another this could be traced. Corrections and erratta in earlier data could be determined quite easily as could new research which obsoletes older research. In the field of genetics this is particularly important(16).

NO ARTIFICIAL SEPARATION OF THE SCIENCES

A special advantage of the SCI is that it can overcome artificial dividing lines that are drawn in various abstracting services in the physical, chemical and

biological sciences. It would be almost impossible to locate a reference to a biological paper made in a physics article by using a conventional abstracting service index. The physics article would be indexed by the Physics Abstracts and the biological paper would be indexed by Biological Abstracts. A search of only one would be incomplete. The inter-disciplinary nature of modern research emphasizes the need for the SCI. A good example of this is the field of instrumentation where one is anxious to know of applications of specific techniques in all branches of science(17).

SCIENTIST TO SCIENTIST COMMUNICATION

A further use of the SCI will be made by the individual scientist interested in determining whether particular lines of research have been pursued by other scientists. Since it is rarely possible to follow up, in the laboratory, all of one's ideas, important data developed by others becomes easily accessible through a citation index. This is especially important to writers of review articles. New areas of research may be proposed as being fertile and the review writer is naturally interested in pursuing the development of these ideas by others.

COMPILING A SCIENCE CITATION INDEX

One of the most attractive features of the SCI is that it is susceptible to almost complete mechanization. Compilation by a staff of trained scientists is not necessary in order to index papers as the "indexing" has already been done by authors in providing citations to earlier papers. Compiling the SCI is almost completely a routine task of copying citations in new papers, sorting them in order by journal, year and page, (so that all references to the same paper will be brought together, and then distributing the information either as a printed bibliography or in card form.

QUANTITATIVE FACTOR PRIMARY CONSIDERATION

The primary factor determining the feasibility of compiling a SCI is a quantitative factor. The first general impression is that there are so many references in the literature as to make a SCI huge and unwieldy. Fortunately this is not true as extensive preliminary studies have shown.

AVERAGE NUMBER OF REFERENCES PER PAPER

In the biological journals the average number of references per paper is fifteen(18). This means that the number of references to the average paper is fifteen. The range of this average is as yet undetermined but one can assume that there may be papers which have never been cited while there are others which are cited hundreds of times.

NUMBER OF PAPERS PER YEAR

The other quantitative factor is the number of papers per year from which one can extract references. 1,000 journals publish approximately 75% of all papers in experimental science. These journals cover every primary field of science. The present coverage of CURRENT CONTENTS is approximately 600 journals which combined publish over 125,000 articles per year. An additional 400 journals in the physical and applied sciences would publish 75,000 more bringing the total to 200,000. Brodman and Taine estimate that 220,000 articles per year should be indexed in a comprehensive index to clinical and experimental medicine(19). If the average experimental paper cites fifteen references we have a total of 3,000,000 citations per year. This is a large number but it is meaningless unless one determines how much effort is required to handle this volume of citations. The effort required must then be compared to the cost and effort required to index 200,000 articles per year by conventional methods. The cost of the average abstract is between six to

ten dollars---about \$2 million per year to which must be added another million or more for indexing the abstracts.

TWO APPROACHES TO THE SCI---COMPREHENSIVE & SELECTIVE

There are basically two approaches that can be taken. If our immediate objective were a complete SCI to all journals then we would obviously have the mechanical task of obtaining, recording, sorting, etc. three million citations per year. This can be done at a cost of between two to three cents per citation---a total of \$60,000 to \$90,000 per year! This surprisingly low figure compares very favorably with the cost of conventional indexing and abstracting. If a "selective" approach is taken one faces the dilemma of establishing the criteria for selectivity. It is fortunate that such criteria can be objectively established for a citation index. Useful results can still be obtained which are not dependent upon any subjective interpretation of what is or is not genetics or any other subject matter.

TIME REQUIRED TO SCAN SELECTIVELY

A study was made of the time required to scan articles for literature references. As many as 200 articles per hour can be scanned even though the "search" criteria varied. Searches were made by criteria such as: author, specific journal, journals whose titles contained the word genetics, or a class of journals such as all general science(GS) journals or journals published by national academies. The cost of obtaining citations "selectively" is naturally higher, per citation found, than taking all references comprehensively. Selective scanning is approximately four times as costly but only 1/16 as many references are processed making the total cost lower. As the number of selection criteria increases and the number of pertinent references found increases one rapidly reaches the point where it is cheaper and more efficient to process every citation. Using the figures above this would be the point at which selective scanning produced one fourth of all references scanned.

WHAT WE PROPOSE TO DO

We propose to construct a citation index to "genetics" based on the selective approach in order to keep the budget of this research project as low as possible without sacrificing all the benefits to be derived from the comprehensive approach. In addition, a lower yearly cost will enable us to process a "backlog" of about five years literature in order that we may effectively demonstrate the value of the SCI. At least this period of time may be needed to bring together enough related material to prove the efficacy of the SCI for up-dating-literature searches.

SELECTION CRITERIA

Our selection criteria are quite simple: (A) All references to a basic list of genetics journals will be included. (B) All references to a basic list of GS journals will be included.

It should be clearly understood that selection of these references will be done simultaneously. The source of these references would be the list of 1,000 journals appended. Scanning will be confined to the last five years of the literature. However, we will process all references made to these journals no matter how old the original article. The product of this project will make it possible to trace current references to very old articles as well as more recent articles.

The rationale for including all references to genetics journals should be readily apparent in compiling a citation index to the field of genetics. However, the reasons for covering the GS journals will be less apparent. Our discussions with geneticists have indicated that any restriction in the project based on a "classical" or "conventional" conception of genetics would be of relatively low value. Some of the most important developments in genetics were reported in

journals which do not contain "genetics" in the title. In addition, a large number of important primary communications in genetics, and all other fields, appear in the GS journals such as Nature and Science. A SCI which covers the GS journals, therefore, would not only cover genetics material but all of its ancillary fields. The pattern of publication for many scientists is a preliminary communication in Science or Nature, followed by a complete paper in some other journal. Through the SCI it will be simple to determine where these complete reports have appeared. Since journals like Experientia, Science, etc. are inter-disciplinary it is also quite possible that any "crossing over" between specialties has its origins in the reading of these journals. Therefore, a SCI to these journals possibly may offer an extremely simple mechanism for quickly permeating the entire scientific literature in any library search.

ONE REFERENCE PER ARTICLE TO A GS JOURNAL

Our studies have shown that on the average there is one reference per article to a GS journal. If we scan the bibliographies and foot notes of 200,000 articles per year we would turn up 200,000 citations per year! (Note: Half of the articles contain no references to GS journals. The other half average two per article giving an overall average of one per article.) At a cost of approximately eight cents per reference the cost of processing 200,000 articles is \$16,000 per year. To cover a five year backlog period the cost is \$80,000. To continue covering the literature for an additional three year period would cost \$50,000. Allowing \$20,000 for other contingencies a budget of \$150,000 for a three year project is proposed. It will be readily apparent that in this type of work the amount of effort can be increased or decreased at will. If our estimates prove to be high or low there will be no difficulty in reducing or increasing journal coverage or the period of time covered. What is important is that the product of this project must be sufficiently complete as to be a fair test of the SCI. The SCI resulting from this project,

fortunately, will be a permanently useful one to every scientist. Since its coverage will be precisely defined scientists and librarians would know exactly what it covered and what it did not. Any use of the SCI could then be supplemented by other conventional approaches. In the future it could be completed without repeating the work already done!

HOW TO PUBLISH THE CITATION INDEX

Critics of the SCI have been alarmed at the potential size of such a compendium. There is really no cause for alarm. Shepard's Citations has been adding over a million references per year and is now in its 85th year! It is not necessary to print all of this information in a single volume even though this too is not impossible through Miniprint(20) or Microcards. Fortunately there is an even simpler solution in science.

SEPARATE CITATION INDEX TO EACH JOURNAL

At the end of this project we will be able to turn over to the editors of the various genetics journals an individual journal citation index to articles published in their journals. Each journal could then publish a yearly supplement consisting of the citation index entries for that journal. Thus, Genetics would publish as a supplement or an article a citation index to articles that have appeared in Genetics. We believe this will also resolve the problem of testing the value of the SCI as all geneticists would have an opportunity to evaluate it through the various society publications.

BUDGET

The budget we have outlined can only be an estimate. In order to minimize the number of permanent employees we have to hire we intend to utilize, if feasible, part-time graduate library students. Since a knowledge of foreign languages can expedite processing, especially the backlog, we intend to explore the feasibility of accepting the generous offer of the Food and Agriculture Organization, Rome to cooperate in this project.(21) More than 60% of the journals required are currently received by the Institute for Scientific Information.

Since it is not efficient to utilize outside library facilities for processing current journals we plan on purchasing some additional journals--those which account for a large percentage of the articles. The balance can be scanned at local libraries. Any additional journals required can be scanned at the various excellent libraries in Philadelphia such as the Franklin Institute Library. We expect to make use of inter-library loans which would also increase journal costs. All backlog material will be scanned by utilizing outside library facilities.

The use of punched card equipment will enable us to sort mechanically and print citation indexes mechanically by use of card operated typewriters or tabulators. Key-punching estimates are based on punching and verifying fifty references per hour. Scanning rates are approximately the same. Since neither job is performed at highest efficiency for long hours we intend to use scanners and key-punchers who can do both jobs alternately. Eight full time scanners and keypunchers could handle all the current material as well as backlog material. One scanner would also handle record keeping. Four key-punching and verifying machines and one sorter should cover our needs adequately at a total equipment cost of \$5,000 per year.

Overhead is figured at 15% of the total budget. \$3,000 per year is allowed toward the salary of the Principal Investigator and an equal amount for a part-time Project Supervisor (Mrs. Gwen Bedford, Univ. Penna. An additional \$4,000 is allowed for travel and other miscellaneous expenses. Travel expenses include the costs of holding one annual advisory board meeting.

Eight scanner-keypunchers.....	\$28,000.00
Project Supervisor.....	3,000.00
Principal Investigator.....	3,000.00
Machine Rentals.....	5,000.00
Travel and Misc.....	4,000.00
Journals & Library facilities...	2,000.00
Overhead.....	<u>7,000.00</u>

Total \$52,000.00

Since the project will extend for a three year period the total three year budget would be \$156,000.

BIBLIOGRAPHY

1. A. H. Seidel, J. Pat. Off. Soc. 31, 554(1949)
2. H. C. Hart, J. Pat. Off. Soc. 31, 714 (1949)
3. E. E. Garfield, J. Pat. Off. Soc. 39, 583-95(1957)
4. W. C. Adair, Am. Document, 6, 31(1955)
5. E. E. Garfield, Science 122, 108-11(1955)
6. E. E. Garfield, Chem. Bull. 43(4), 11-12(1956)
7. Shepard's Citations, Inc., Colorado Springs, Colo.
8. Hearings, House of Representatives, Comm. on Approp., Feb. 19, 1957, p. 1392
9. M. M. Berry, Aug. 22, 1957
10. Eugene Garfield Associates, "General Feasibility Study of Citation Indexes for Science". A proposal for Research, June 1958
11. D. E. Gray, Oct. 23, 1958
12. G. Allen, private communication, Jan. 24, 1957
13. J. Lederberg, private communication, May 9, 1959
14. G. Lefevre, formerly NSF Genetics Section
15. K. S. Wilson, Exec. Secy, Genetics Study Section, NIH & Dr. Green, Bar Harbor, Maine
16. G. Allen, letter to B. W. Adkinson, NSF, Sept. 5, 1958
17. J. Stern, private communication
18. E. E. Garfield, letter to J. Lederberg, Sept. 9, 1959
19. E. Brodman & E. Taine, Int'l. Conf. Sci. Inform. Proc. Vol.I, 435-47(1958)
20. E. E. Garfield, ibid. 461-74(1958) p.465
21. G. L. Kesteven, FAO, Fisheries Biology Branch, Rome, Oct.10, 1958 & June 4,1959